USING THE LITERATURE BASED DISCOVERY RESEARCH METHOD IN A CONTEXT OF BUILT ENVIRONMENT RESEARCH

Nathan Kibwami[1] and Apollo Tutesigensi

[1] Institute for Resilient Infrastructure, School of Civil Engineering, University of Leeds, Leeds, LS2 9JT, UK

For two disparate research groups, unaware of each other's work, one group can inadvertently solve a problem prevalent in the other. Without considering work from both groups together, such breakthroughs may remain undiscovered. The solution is literature based discovery (LBD), a method which involves investigation or search for novel hypotheses connecting work from two or more disparate contexts. However, LBD has predominantly been used to address medical problems, and its uptake outside medical research remains scanty. In the context of built environment research, there are countable studies that have claimed using LBD and moreover, they presented sparse details. On one hand, studies that have claimed using LBD as a research method seem to confuse it with traditional literature reviews, and on the other hand, even those that could have used LBD seem unaware that they used some kind of LBD-style analysis. Following the original principles of LBD, this paper presents an LBD-inspired research method and a demonstration of its applicability within a built environment research context. The findings indicate promising leads to encouraging LBD and elucidating several misconceptions surrounding its use in built environment research. It is hoped that this paper will encourage future research in built environment, like construction management research, to confidently use LBD appropriately and consciously.

Keywords: built environment, carbon emissions, literature based discovery.

INTRODUCTION

The objective of any piece of research is to advance knowledge within the respective field or context of inquiry. There is a possibility, however, that a problem prevalent in certain field or context might be unknowingly solved by another disparate field, oblivious to the problem (Hristovski *et al*. 2005). Such inadvertent breakthroughs can remain undiscovered and consequently unpursued, if no inquiry ever considers the disparate fields, together. Revelation of such undiscovered knowledge is the domain of literature based discovery (LBD). LBD is a form of text interrogation of scientific literature to identify *"... nontrivial assertions that are implicit, and not explicitly stated ..."* (Smalheiser 2012: 218). It is argued that coalescing disjoint literature can generate hypotheses and subsequently, yield new knowledge (Weeber *et al*. 2001).

Coined in biomedicine-related works (Swanson and Smalheiser 1997; Swanson 1986), LBD has proliferated through several studies (Smalheiser 2012; Smalheiser *et al*. 2009; Torvik and Smalheiser 2007; Srinivasan 2004), but a few (e.g. Ittipanuvat *et al*.

[1] cnnk@leeds.ac.uk

Kibwami, N and Tutesigensi, A (2014) Using the literature based discovery research method in a context of built Environment research *In:* Raiden, A B and Aboagye-Nimo, E (Eds) *Procs 30th Annual ARCOM Conference, 1-3* September 2014, Portsmouth, UK, Association of Researchers in Construction Management, 227-236.

2013; Yung *et al*. 2013; Dixit *et al*. 2010; Kostoff *et al*. 2008b) outside medical research. In the context of the built environment research, LBD was used to identify parameters that led to variations and inconsistences in buildings' embodied energy (Dixit *et al*. 2010). In another paper (Dixit *et al*. 2013), it is reported that a model meant to facilitate lifecycle energy analyses of buildings was developed aided by the LBD method. Yung also acknowledged using LBD to analyse several previous studies related to energy analyses of buildings (Yung *et al*. 2013). However, there is limited articulation of why and how the LBD method was used, since none of the studies (Dixit *et al*. 2010; Dixit *et al*. 2013; Yung *et al*. 2013) provided any details beyond citations. LBD uptake outside medical research, especially in built environment research, remains promising but sparsely articulated.

There are several studies that can be placed in the context of the built environment (e.g. Egebjerg 2013; Ibn-Mohammed *et al*. 2013; Dixit *et al*. 2012; Dakwale and Ralegaonkar 2011; Ramesh *et al*. 2010; Sartori and Hestnes 2007; Casals 2006) for which there is sufficient evidence to suggest that LBD could have been an appropriate method to use, or at least cite. Analysing each of these studies is beyond the scope of this paper but attention is drawn to a particular one. Based on literature and interviews, the study (Egebjerg 2013) compared the movie industry (i.e. film making) and the construction industry, with the aim of identifying what the latter could learn from the former. Such a methodology is reminiscent of LBD, where LBD hypotheses (i.e. how movies can inspire construction) are generated and then corroborated through empirical studies. For instance, Swanson's LBD hypotheses (Swanson 1986) were experimentally corroborated in DiGiacomo *et al*. (1989). For Egebjerg (2013), both LBD-hypothesising and corroboration could have been possible. On the whole, the prevailing mysteries regarding use of LBD in built environment research raise intriguing questions worth to explore. In concurrence with assertions in Smalheiser (2012), it is likely that researchers never recognise the difference between carrying out an LBD analysis and the traditional literature/text analyses, while others seem unaware of LBD.

The foregoing discussions raise several concerns regarding the efficacy of LBD adoption in built environment research:

- no study exists to underscore the efficacy of LBD method in built environment research;
- LBD has previously been inappropriately or redundantly cited as the research method, moreover with insufficient details provided;
- it is insufficient for a study to claim using LBD just because it based its analysis and findings on literature;
- LBD could have been the appropriate method for many studies that did not mention any method used; and
- it is likely that some researchers are unaware that they used some kind of LBD-style analysis and such ignorant use is unlikely to be efficient.

While these concerns are rather many to exhaustively address in a single writing, work presented in this paper shades light on several of them by presenting an LBD method and demonstrating its application to a particular case (i.e. addressing carbon emissions from buildings) that falls in the context of sustainability in the built environment.

A BRIEF REVIEW OF LBD

The founding works in LBD were based on the ‘ABC’ approach, popularised by a Venn diagram (see Swanson and Smalheiser 1997: 184). This form of syllogism prescribes that for mutually isolated literatures A and C, if A reports a relationship (AB) with a term/concept B, C reports a relationship (BC) with the same term/concept B, hypotheses (AC) can be derived connecting A and C (Smalheiser 2012). If nothing has previously been explicitly reported regarding the connection between A and C, then it can be of significant interest or rather, a new form of discovery (Weeber *et al*. 2001; Swanson and Smalheiser 1997). From two scientific literatures, one related to fish oil and another on Raynaud's disease (a condition of intermittent blood flow), Swanson *"proposed [a] hypothesis that fish oil might ameliorate Raynaud's syndrome"* (Swanson 1986: 12). Considering blood viscosity (term/concept B), patients with Raynaud's disease (Literature A) had abnormalities in blood viscosity (AB), yet eicosapentaenoic acid (Literature C) found in fish oils lowers blood viscosity (BC). Considering literature A and C, a connection AC was hypothesised that patients of Raynaud's disease could benefit from Fish oils (Swanson and Smalheiser 1997; Swanson 1986). Several subsequent LBD studies were undertaken some of which were successfully corroborated (DiGiacomo *et al*. 1989) or replicated (Srinivasan 2004; Weeber *et al*. 2001; Lindsay and Gordon 1999). A predominant LBD process is that denoted as a closed discovery (CD) approach (Kostoff *et al*. 2008a; Weeber *et al*. 2001). The CD process simultaneously starts with the disease (C) and fish oils (A), looking for common linking B-terms like blood viscosity, working towards identifying linkages between A and C (Kostoff *et al*. 2008a). The CD process resonates with LBD in form of two-node analyses and indeed, some researchers (Smalheiser *et al*. 2009) have gone ahead to develop tools (Arrowsmith 2007) that can guide carrying out CD LBD processes. Unfortunately, all these efforts are still limited to medical research.

THE PROPOSED LBD RESEARCH METHOD

Following the CD process, the proposed LBD method can be summarised into two phases, composed of six stages (see Table 1).

First phase (stages 1 to 4)

A comprehensive literature search is performed on two disparate 'contexts' of inquiry (i.e. A and C) to generate the corpora (i.e. peer reviewed journal articles) for performing LBD. Using adequate linguistic specifications (see Frantzi *et al*. 1998) and appropriate software, terms are automatically extracted from each context. 'Recall' and 'Precision' are two aspects that the Term extraction process has to optimise (Naumann and Herschel 2010). 'Recall' relates to the number of terms that can be retrieved, whereas 'precision' is related to the relevance or plausibility of the extracted terms. Higher 'precision' can only be guaranteed at the expense of lower 'recall', and vice versa (Ganti and Sarma 2013).

In well-structured and online corpora (e.g. in MEDLINE), it is possible to know the approximate number of terms to work with (see Weeber *et al*. 2001: 551). However, for a semi-automated process suggested in this work (i.e. articles manually gathered from different databases), only an estimate can be possible. For instance, for literature consisting 20 articles, assuming an average full-article length of 7000 words, this would constitute working with 140,000 terms. To manage the winnowing process towards precision, an initial working number of terms from each context should be set. Meanwhile, the decision of setting the minimum term length (number of characters per term) depends on the desired precision and recall. Shorter terms are usually good on recall but not precision. Also, terms can be unigrams (i.e. one word terms), bigrams (i.e. two word terms) or n-grams (Frantzi *et al*. 1998; Ittipanuvat *et al*.

2013). Because of some limitations highlighted later, the current approach considers unigrams. The ‘recall’ for unigrams is usually high since, unigrams can exist either on their own, or as nested terms (i.e. sub-terms of bigrams/n-grams). In Ittipanuvat *et al.* (2013), unigrams accounted for over three quarters of the total terms extracted. Though unigrams present rather a bigger sample space to work with, measures (e.g. a stop word list) need to be undertaken in order to ensure precision (i.e. limiting retrieval of irrelevant terms).

The stop word list is primarily a tool used to distinguish between potentially useful and non-useful terms (e.g. frequently occurring terms like ‘is’, ‘to’, ‘what’ etc.) (Weeber *et al*. 2001) and it has previously been employed in LBD studies (see Lindsay and Gordon 1999; Swanson and Smalheiser 1997). The stop list can be precompiled based on the predicted suitability of terms (Swanson and Smalheiser 1997) or compiled concurrently with the term extraction process (Lindsay and Gordon 1999). As part of the synonymy and stemming rules (Lindsay and Gordon 1999), it is suggested that only exactly (i.e. not synonyms) matching words should be considered in order to control unnecessary recall and noise. However, singular-plural stemming rules (Lindsay and Gordon 1999) can be applied and in such cases, the terms (e.g. house and houses) should be combined into one. Illustrations of linguistic filtering reveal that most terms are usually composed of nouns, verbs or adjectives and for multiword terms, they are usually constituted of at least a noun (Frantzi *et al*. 1998). Thus terms extracted should be linguistically filtered to nouns, verbs and adjectives in that order of preference. For automated filters like those in medical databases, linguistic filtering can be automatically set. The approach suggested herein is semi-automated since linguistic filtering is manually done by inspecting each extracted term.

Lexical statics are used to manipulate the extracted lists of terms in order to retain the most plausible ones. Initially, terms should be sorted/ranked by their frequency (Tf) (i.e. number of times a term appears in the corpus). However, using term frequency alone for further evaluations means that terms appearing less frequently might be missed out (i.e. since they are low ranked), yet they may be plausible. To circumvent this, the concept of inverse document frequency (iDf) developed in Jones (1972) can be used. The iDf weighting boosts terms with low frequency, yet concentrated in few specific documents/articles. This consequently yields a Term frequency-Inverse document frequency (Tf-iDf) measure (Salton and Buckley 1988). Tf-iDf (see formula in Table 1 footnote) is a more preferred measure and has been cited in LBD studies (Lindsay and Gordon 1999; Ittipanuvat *et al*. 2013; Srinivasan 2004) as a better measure of relevance of a term than frequency alone. Therefore, terms should be ranked by Tf-iDf and low-ranking terms may be discarded. Unlike in biomedical databases where terms can be automatically classified into their respective predetermined semantic categories (see Smalheiser *et al*. 2009), manual categorisation is suggested, which rather demands 'human intervention' and acquaintance with qualitative data analysis techniques. However, this does not entirely manifest as a disadvantage since it gets the analyst up-close with ‘what the literature is saying’. To guide the categorisation process, a paradigm model, initially proposed in Strauss and Corbin (1998) and subsequent texts (Corbin and Strauss 2008), is suggested. It consists of Phenomena (i.e. what is going on?), Conditions (i.e. what are the causes), Actions/interactions (i.e. what is the response?) and Consequences (i.e. what are the results?). Categories are developed (i.e. using appropriate software) from key terms and it is possible for a given term to belong in several categories. Essentially, a term is evaluated (according to where it appears in the text) and the approach involves ‘coding’, *“… an analytic process through which data are fractured, conceptualised, and integrated …”* (Strauss and Corbin 1998: 3). Coding should be done by sentence and paragraph, through questioning the major idea embedded therein.

Table 1: The proposed LBD method

No.	Step	Procedures	Remarks
1	Literature data retrieval	Specify context/dichotomy (i.e. A and C). Identify core /key terms. Identify literature databases. Retrieve literature for each dichotomy in relation to key terms.	This is where the boundary of literature is specified. Appropriate search arguments are crucial to ensure comprehensiveness.
2	Term extraction and linguistic specification	Specify the number of terms and minimum characters per term. Specify number of words/strings per term. Specify level of term synonymy. Specify desired linguistic filter. Build a stop word list.	This demands a balance between 'recall' and 'precision'. Appropriate software can be used to extract terms from the texts/articles.
3	Lexical statistics*	Retrieve Term frequency (Tf). Retrieve document number (D). Retrieve Document frequency (Df). Compute Term frequency-inverse document frequency (Tf-iDf). Rank terms to identify most plausible ones.	Tf, D and Df values are computed by the software used in extraction. Terms are ranked by Tf-iDf and an appropriate cut-off list is selected.
4	Category development	Code around key terms using the paradigm model (i.e. Phenomena, Conditions, Actions/interactions' and Consequences). Identify major categories that emerge.	Coding is done using software. Key terms are coded by sentence and paragraph. Other terms are coded to where they appear in text. Categories are developed by assessing the context of how the term is used.
5	Semantic similarity	Compute Cosine similarity of vectors/categories for A and C.	Query terms by context per vector. Compute similarity of vectors.
6	Deducing relations	Investigate similarity of vectors and identify top ranked terms per vector. Make inferences.	If a vector in A is similar to that in B, then the terms contained can be related by a hypothesis

* Tf is the number of times a term appears, D is the total number of documents, Df is the number of documents in which the term appears and Tf-iDf is computed as Tf x log(D/Df).

Second phase (stages 5 and 6)

Literature (Ganti and Sarma 2013; Naumann and Herschel 2010; Pesquita *et al*. 2009) discusses several similarity measures (e.g. Jaccard Index, Dice coefficient and cosine) which are also often used in LBD studies (see Ittipanuvat *et al*. 2013; Miyanishi *et al*. 2010). This work considers the cosine similarity measure. At this stage, categories based on the key terms only would have been developed from the corpora. From appropriate coding software, it is possible to retrieve terms that intersect with a given category; several term-combinations (e.g. terms in both A and C, in A only and in C only) associated with a developed category can be worked out. It is possible for a given term to belong in several categories. The categories are transformed into vectors for which similarity between them is computed. Since vectors only work with integers, each term is therefore represented by its Tf-iDf measure. Put another way, a category composed of terms is represented as a vector composed of Tf-iDfs. This idea, initially suggested in Salton and Buckley (1988), is usually used in works related to document indexing and retrieval. The similarity between two vectors is a property of the cosine of the angle between them (i.e. 1 if the vectors are identical and 0 if they are not). The cosine values are computed using the cosine vector similarity formula (Salton and Buckley 1988: 514) as per the Equation below:

$$Similarity\ (A_v, C_v) = \sum\left(w_{A,t} \times w_{C,t}\right) / \left(\sqrt{\sum w_{A,t}^2} \times \sqrt{\sum w_{C,t}^2}\right)$$

where A_v and C_v are Tf-iDf vectors representing literature contexts of A and C respectively; $w_{A,t}$ and $w_{C,t}$ are the weights (i.e. Tf-iDf) of a term t with regard to literature A and C respectively.

The deduction of relationships (i.e. plausible hypotheses) is then based on the cosine similarity measure and the Tf-iDf measure. It is assumed that vectors (i.e. categories) with cosine similarity values closer to 1 will be more related and thus plausible for generating plausible hypotheses. This assumption is rather not new (see Miyanishi *et al*. 2010: 1554), though needs cautious interpretation. Although it would be considered that the lower cosine values offer fewer linkages to explore, they may, ipso facto, be potential sources for novel relationships. Nonetheless, the key guidance to pursue any plausible hypothesis/relationship regarding any term in the vectors is based on the cosine similarity score and the term's rank/weighting (i.e. Tf-iDf). In other words, it is inferred that the plausibility of a hypothesis linking an A-Term to a C-Term is related to the cosine similarity between the two vectors that describe how that term manifests in A and C.

APPLICATION OF THE METHOD

Personal experience and anecdotal evidence suggested that there was little, if anything, implemented in the Ugandan building sector to address carbon emissions. This was confirmed by a nil return when a systematic search for literature involving the key words of 'building(s)' or 'construction' and 'carbon emissions' was implemented. A similar search involving the United Kingdom (UK) returned a rich collection of publications, suggesting that there may be lessons that can be learnt from the UK for the Ugandan context. Therefore, considering carbon emissions in the building sector in Uganda as an area for investigation in the context of promoting sustainable construction was warranted. A comprehensive literature search was performed to retrieve literature from Uganda (denoted as A) and that from UK (denoted as C). The searched databases (in fields of Title, Abstract and Key words) considered to cover most of the academic journals in English, were Science Direct, Engineering Village, ProQuest, EBSCO Host and Web of Knowledge/Science. The key words considered for A were Uganda and emissions, whereas for B, UK or United Kingdom, buildings or construction, and emissions. A typical search argument was in form of: search in Title-Abstract-Key words (United Kingdom OR UK) AND in Title-Abstract-Key words (emissions OR greenhouse gases) AND in Title-Abstract-Key words (building* OR construct*). After filtering, a total of 105 articles were considered, 29 for A and 76 for C. The articles were imported into Nvivo 10 software (see Bazeley and Jackson 2013; QSR Nvivo 2013) to extract terms. Consequently, appropriate linguistic filters and a stop word list were applied in term extraction. While there was no limit on the maximum number of characters in a term (i.e. Term length), a minimum was arbitrarily set to three. Although Nvivo 10 comes with a default stop word list, more terms were progressively added, and the total came to around 8000 words. To balance precision and recall, while presenting a manageable number of terms, only 1000 terms were extracted from each context A and C. Lexical statics (i.e. Tf-iDf) were computed and used to rank the extracted terms. This was performed by exporting the lists to Microsoft Excel and applying appropriate formulae. Each of the extracted terms was then converted into a node (i.e. Nvivo10 nodes function) in order to 'tether' it with the document/article (and precise location) it appeared in. The name of the node corresponded to the name of the term. Using the paradigm model as described before, categories were developed (i.e. using Nvivo10 coding) around the key search terms (i.e. emissions and buildings).

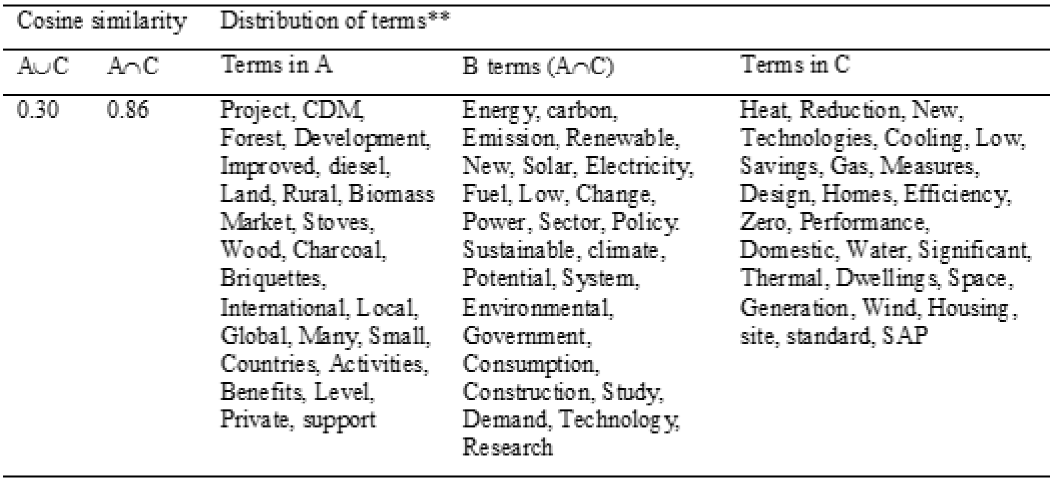
Using Nvivo 10 query functions (i.e. Matrix coding queries), it was possible to generate terms that belonged to each of the developed categories, with respect to A and C, ranked by their Tf-iDf measure. For each pair (i.e. a given category/vector but split into terms as mapped to A, and terms as mapped to C), the cosine similarity value was computed. For instance, a vector (representing a category named 'strategies to reduce emissions') was mapped to A, and the resulting list of terms noted/ranked. The same vector was then mapped to C and also the resulting list of terms noted/ranked. The cosine similarity of the two vector's lists was then computed. If the similarity was found to be (or nearly) zero it implied that the 'strategies to reduce emissions' in A were (almost) 100% dissimilar to those in C. The resulting terms, using appropriate sorting facilities in Microsoft Excel, were then investigated to isolate: terms in the vector that are present in both A and C ($A \cap C$), A only, C only and the union of A and C ($A \cup C$). In arriving at relationships, the terms were validated by re-examining the article where they appeared, in line with the vector's description.

FINDINGS AND DISCUSSIONS

The top ten terms (out of 1000) considered in A showed that: biomass, forests, urban, CDM (clean development mechanism), electricity, charcoal, wood, stoves, solar and land are important terms in relation to emissions, whereas in C, cooling, SAP (standard assessment procedure), embodied, dwellings, housing, wind, electricity, stock, zero and office are important terms in relation to emissions and buildings. Electricity, solar, power, climate, renewable, energy, costs, construction, fuel and technology were the top ten terms present in both A and C, ranked by their average normalised Tf-iDf scores. In LBD nomenclature, these are referred to as the B-terms. They are the ones that create the primary linkages between A and C. So, the relationship (AB) of such a B term with respect to A, compared with the relationship (BC) of the same term but with respect to C, forms the primary basis of LBD analyses. Some of such relationships can be the same (e.g. if a 'causes of emissions' vector returns 'energy' in both A and C), justifying that an issue is indeed prevalent in both A and C, while other relationships might differ. The discretion of what to investigate or pursue then lies in the hands of the analyst, guided by the aim of carrying out the LBD study.

Several categories were developed and the similarity in relation to B terms ranged from 0.81 to 0.86, with the 'strategies to reduce emissions' category scoring highest. Since the maximum score was 1, the results implied a good relationship between A and C, a signal that perhaps similar actions exist in both contexts. When terms not common to both contexts were also added (i.e. A∪C), the similarity reduced i.e. by 48% in causes of emissions, 43% in barriers to reducing emissions, 65% in strategies to reducing emissions and 44% in regulations related to emissions. It was clear that there was a significant difference between A and C, regarding the strategies to reducing emissions, compared with other categories. From such information (see Table 2), several relations could be deduced and once corroborated, yield new knowledge. For instance, an assertion can be posed that CDM (i.e. a highly ranked term in A) can be used to address emissions from buildings through applying SAP (i.e. a highly ranked term in C). Certainly, literature searches performed indeed proved that there was no current research to support that assertion, and this could be a potential research problem for which once empirical research is conducted, yield new knowledge. To the discretion of the analyst, a plethora of plausible suppositions can be elicited from the information.

Table 2: Similarity with respect to 'strategies to reduce emissions' category



Cosine similarity		Distribution of terms**		
A∪C	A∩C	Terms in A	B terms (A∩C)	Terms in C
0.30	0.86	Project, CDM, Forest, Development, Improved, diesel, Land, Rural, Biomass Market, Stoves, Wood, Charcoal, Briquettes, International, Local, Global, Many, Small, Countries, Activities, Benefits, Level, Private, support	Energy, carbon, Emission, Renewable, New, Solar, Electricity, Fuel, Low, Change, Power, Sector, Policy. Sustainable, climate, Potential, System, Environmental, Government, Consumption, Construction, Study, Demand, Technology, Research	Heat, Reduction, New, Technologies, Cooling, Low, Savings, Gas, Measures, Design, Homes, Efficiency, Zero, Performance, Domestic, Water, Significant, Thermal, Dwellings, Space, Generation, Wind, Housing, site, standard, SAP

** In each column terms are listed separated by commas, beginning from the highest as ranked by Tf-iDf weighting.

CONCLUSIONS

This paper has highlighted the efficacy of the literature based discovery (LBD) method in built environment research. Although LBD has proliferated, it is mostly still limited to addressing medical problems and uptake in built environment research is promising but sparsely articulated. Built environment research seems to confuse the difference between carrying out LBD analyses and the traditional literature searches/reviews, a situation that culminates into stifling efforts to proliferate LBD. An LBD research method was presented and a demonstration of its application in the context of built environment research provided. In the example, a relationship between carbon emissions, clean development mechanism (CDM) and the standard assessment procedure (SAP) of buildings was coined. The overall findings underscore the untapped potential of LBD in built environment research and moreover, LBD adds scientific rigour to traditional literature review techniques. However, some limitations were met. Unlike biomedical research, the approach used was semi-automated and thus some tasks were limited to only what could be reasonably and manually handled.

In addition, the LBD analysis demanded some levels of subjectivity since terminology and databases related to built environment research are not as highly specific and descriptive as those of medical research. Although software was used, it was not tailor-made for LBD. For instance, Nvivo10 was not able to automatically extract multiword terms. Nevertheless, LBD remains a potentially appropriate research method that can address problems in built environment. It is hoped that more work will build on this information to encourage use of LBD in wider fields of built environment.

REFERENCES

Arrowsmith (2007) *"Arrowsmith Project: linking documents, disciplines, investigators and databases."*. [Online]. [Accessed November 10, 2013]. Available from: http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html.

Bazeley, P and Jackson, K (2013) *"Qualitative data analysis with NVivo"*. 2nd ed. London: SAGE.

Casals, X G (2006) Analysis of building energy regulation and certification in Europe: Their role, limitations and differences. *"Energy and Buildings"*, **38**(5), 381-392.

Corbin, J M and Strauss, A L (2008) *"Basics of qualitative research : techniques and procedures for developing grounded theory"*. 3rd ed. Los Angeles, Calif. ; London: SAGE.

Dakwale, V A and Ralegaonkar, R V (2011) Review of carbon emission through buildings: threats, causes and solution. *"International Journal of Low-Carbon Technologies"*, **7**(2), 143-148.

DiGiacomo, R A, Kremer, J M and Shah, D M (1989) Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. *"American Journal of Medicine"*, **86**, 158-64.

Dixit, M K, Culp, C H and Fernández-Solís, J L (2013) System boundary for embodied energy in buildings: A conceptual model for definition. *"Renewable and Sustainable Energy Reviews"*, **21**, 153-164.

Dixit, M K *et al*. (2010) Identification of parameters for embodied energy measurement: A literature review. *"Energy and Buildings"*, **42**(8), 1238-1247.

Dixit, M K *et al*. (2012) Need for an embodied energy measurement protocol for buildings: A review paper. *"Renewable and Sustainable Energy Reviews"*, **16**(6), 3730-3743.

Egebjerg, C (2013) Learning from movie-sets coordination. In: S.D. Smith and D.D. Ahiaga-Dagbui, (Eds), *"29th Annual Conference on Association of Researchers in Construction Management, ARCOM 2013, "* 2-4 September 2013, Reading, UK. Association of Researchers in Construction Management (ARCOM), **2**, 827-836.

Frantzi, K, Ananiadou, S and Tsujii, J (1998) The C-value/NC-value Method of Automatic Recognition for Multi-word Terms. *"Research and Advanced Technology for Digital Libraries"*. Springer Berlin Heidelberg.

Ganti, V and Sarma, A D (2013) Data Cleaning: A Practical Perspective. *"Synthesis Lectures on Data Management"*, **5**(3), 1-85.

Hristovski, D *et al*. (2005) Using literature-based discovery to identify disease candidate genes. *"Int J Med Inform"*, **74**(2-4), 289-98.

Ibn-Mohammed, T *et al*. (2013) Operational vs. embodied emissions in buildings—A review of current trends. *"Energy and Buildings"*, **66**, 232-245.

Ittipanuvat, V *et al*. (2013) Finding linkage between technology and social issue: A Literature Based Discovery approach. *"Journal of Engineering and Technology Management"*, http://dx.doi.org/10.1016/j.jengtecman.2013.05.006.

Jones, K S (1972) A statistical interpretation of term specificity and its application in retrieval. *"Journal of Documentation"*, **28**(1), 11-21.

Kostoff, R N *et al*. (2008a) Literature-related discovery (LRD): Methodology. *"Technological Forecasting and Social Change"*, **75**(2), 186-202.

Kostoff, R N *et al*. (2008b) Literature-related discovery (LRD): Water purification. *"Technological Forecasting and Social Change"*, **75**(2), 256-275.

Lindsay, R K and Gordon, M D (1999) Literature-based discovery by lexical statistics. *"Journal of the American Society for Information Science"*, **50**(7), 574-587.

Miyanishi, T, Seki, K and Uehara, K (2010) Hypothesis generation and ranking based on event similarities. In: *"Proceedings of the 2010 ACM Symposium on Applied Computing"*, Sierre, Switzerland. 1774421: ACM, 1552-1558.

Naumann, F and Herschel, M (2010) An Introduction to Duplicate Detection. *"Synthesis Lectures on Data Management"*, **2**(1), 1-87.

Pesquita, C *et al*. (2009) Semantic Similarity in Biomedical Ontologies. *"PLOS Computational Biology"*, **5**(7).

QSR Nvivo (2013) *"Nvivo10 for Windows help"*. [Online]. [Accessed December 18, 2013]. Available from: http://webhelp-nv10.qsrinternational.com/nv10_help.htm.

Ramesh, T, Prakash, R and Shukla, K K (2010) Life cycle energy analysis of buildings: An overview. *"Energy and Buildings"*, **42**(10), 1592-1600.

Salton, G and Buckley, C (1988) Term-weighting approaches in automatic text retrieaval. *"Information Processing and management"*, **24**(5), 513-523.

Sartori, I and Hestnes, A G (2007) Energy use in the life cycle of conventional and low-energy buildings: A review article. *"Energy and Buildings"*, **39**(3), 249-257.

Smalheiser, N R (2012) Literature-based discovery: Beyond the ABCs. *"Journal of the American Society for Information Science and Technology"*, **63**(2), 218-224.

Smalheiser, N R, Torvik, V I and Zhou, W (2009) Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *"Computer Methods and Programs in Biomedicine"*, **94**(2), 190-197.

Srinivasan, P (2004) Text mining: Generating hypotheses from MEDLINE. *"Journal of the American Society for Information Science and Technology"*, **55**(5), 396-413.

Strauss, A L and Corbin, J M (1998) *"Basics of qualitative research : techniques and procedures for developing grounded theory"*. 2nd ed. Thousand Oaks: Sage Publications.

Swanson, D R (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *"Perspectives in Biology and Medicine"*, **30**(1), 7-18.

Swanson, D R and Smalheiser, N R (1997) An interactive system for finding complementary literatures: A stimulus to scientific discovery. *"Artificial Intelligence"*, **91**(2), 183-203.

Torvik, V I and Smalheiser, N R (2007) A quantitative model for linking two disparate sets of articles in MEDLINE. *"Bioinformatics"*, **23**(13), 1658-1665.

Weeber, M *et al*. (2001) Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *"Journal of the American Society for Information Science and Technology"*, **52**(7), 548-557.

Yung, P, Lam, K C and Yu, C (2013) An audit of life cycle energy analyses of buildings. *"Habitat International"*, **39**, 43-54.